Q1:

PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

Q2:

Survived, Pclass, Sex, SibSp, Parch, Fare, Embarked

Q3:

Age, SibSp, Parch, Fare

Q4:

Ticket, Cabin

Q5:

In training set: Age, Cabin, Embarked

In test set: Age, Cabin

Q6:

PassengerId: integer

Survived: integer

Pclass: integer

Name: string

Sex: string

Age: float

SibSp: integer

Parch: integer

Ticket: string

Fare: float

Cabin: string

Embarked: character

Q7:

Age

count 714

mean 29.69911764705882

std 14.516321150817316

```
min  0.42
25% percentile 20.125
50% percentile 28.0
75% percentile 38.0
max  80.0
SibSp
count  891
mean  0.5230078563411896
std  1.1021244350892878
min  0
25% percentile 0.0
50% percentile 0.0
75% percentile 1.0
max  8
Parch
count  891
mean  0.38159371492704824
std  0.8056047612452208
min  0
25% percentile 0.0
50% percentile 0.0
75% percentile 0.0
max  6
Fare
count  891
mean  32.2042079685746
std  49.6655344447741
min  0.0
25% percentile 7.9104
```

50% percentile 14.4542

75% percentile 31.0

max 512.3292

Q8

Survived

count 891

unique 2

top

[0]

freq

[549]

Pclass

count 891

unique 3

top

[3]

freq

[491]

Sex

count 891

unique 2

top

['male']

freq

[577]

```
SibSp
count 891
unique 7
top
 [0]
freq
 [608]

Parch
count 891
unique 7
top
 [0]
freq
 [678]

Fare
count 891
unique 248
top
 [8.05]
freq
 [43]

Embarked
count 889
unique 3
top
 ['S']
```

freq

[644]

Q9

correlation is -0.33848103596101514

Not significant. I will not include this feature in the predictive model

Q10

We use male=0 female=1 and calculate the correlation between sex and survived.

the correlation is 0.5433513806577546, which is significant. Since the number is negative, which means female is more likely to have survived.

Q11

Do infants (Age <=4) have high survival rate?

Yes, they do.

Do oldest passengers (Age = 80) survive?

Yes, they do.

Do large number of 15-25 year olds not survive?

Large number of 15-25 year old people do not survive.

Based on your analysis of the histograms,

Should we consider Age in our model training? (If yes, then we should complete the

Age feature for null values.)

Yes.

Should we should band age groups?

Yes.

Q12

Does Pclass=3 have most passengers, however most did not survive?

Yes.

Do infant passengers in Pclass=2 and Pclass=3 mostly survive?

Yes.

Do most passengers in Pclass=1 survive?

Yes.

Does Pclass vary in terms of Age distribution of passengers?

Yes. Older people with higher Pclass.

Should we consider Pclass for model training?

Yes.

Q13

Do higher fare paying passengers have better survival?

Yes.

Should we consider banding fare feature?

Yes, because so many tickets have the same fare.

Q14

What is the rate of duplicates for the Ticket feature?

0.2356902356902357

Is there a correlation between Ticket and survival?

Yes.

Should we drop the Ticket feature?

Yes, since we would like to include Pclass and the ticket number is highly related to Pclass, we do not need to include Ticket feature.

Q15

Is the Cabin feature complete?

No.

How many null values there are in the Cabin features of the combined dataset of training and test dataset?

1014

Should we drop the Cabin feature?

No. We have so many null data, which means we cannot find ways to make up.

Q16

Please check my python codes

Q17

I do not see any questions

Q18

Based on the results from Q8, we should use 'S' for missing values

Q19

We use 7.75 for missing values

Q20

Please check my python code.

Approximately how many hours did you spend on this assignment?

5 hours

Which aspects of this assignment did you find most challenging?

Plot the data.

Were there any significant stumbling blocks?

The provided examples is not very clear, but with professor's clarifications, everything is smooth.

Which aspects of this assignment did you like?

Show us how to deal with diffrent data problems (data missing, feature selection...)

Is there anything you would have changed?

I would say we could add some links for panda library tutorials.